疑似乱数検定のための直交検定系 (確率数値解析に於ける諸問題, Ⅳ)

AUTHOR(S):

杉田, 洋

CITATION:

杉田, 洋. 疑似乱数検定のための直交検定系 (確率数値解析に於ける諸問題, Ⅳ). 数理解析研究所講究録 2000, 1127: 80-87

ISSUE DATE:

2000-01

URL:

http://hdl.handle.net/2433/63611

数理解析研究所講究録
1127巻 2000年 80-87

疑似乱数検定のための直交検定系

杉田　洋　(九州大学大学院数理学研究科)

序

疑似乱数はランダムに選ばれる「種」から決定論的アルゴリズムに従って生成される確率過程である．問題は種の小さなランダム性を引き延ばして，大きな（長い）ランダム列に見せ掛けることである．それがうまく行ったかどうかは，多種の「検定」を行って総合的に判断することになる．そのとき，どんな検定を行えばよいのだろうか．

教科書等(e.g.[1])にある検定法を眺めていると，どうもあまり組織的でないような印象を受ける．とくに，種々の検定法の相関関係が明瞭でない．たとえば検定法A と検定法B の間に強い相関がある場合は，基本的にどちらか一方を実施すればよく，多くの手間を費やして両方とも実施する必要はない．その時間をまったく相関のない検定法C に費やすのがよかろう．従って，各検定法の相関を予め調べておくことが望ましいだろうが，それは容易なことではない．

そこで，初めから相関の様子が分かっている検定法たちを用意しておくのはよいことだろう．ここでは，どの二つの検定も無相関であるような検定系，—「直交検定系」と呼ぶ—，について調べる．直交検定系は系全体としては無相関ではないけれども，それに近い性質を持つ．

直交検定系の例として種々の「パリティ検定系」を提唱する．パリティ検定系はただ1つの検定法ではなく，あるパラメータ U を持つ直交する検定法の集合体である．パラメータ U の取り方は，事実上，無数にあるので，相関の小い無数の検定を行うことができる．まだ，完全に組織的な数値実験には至ってないが，メルセンヌ・ツイスター(最大周期列法 — いわゆるM系列 — の一種)に関する検定結果を例として挙げた．

最近，暗号関係の疑似乱数に関連して，それらを破ろうとする大規模な試みが主にネットワークを通じて行われている．そして，安全性の高いと思われていた手法のいくつかが実際に破られている．翻って，数値計算用の疑似乱数の場合，実際に行われている検定は徹底性を欠いており，優秀と言われている疑似乱数を徹底的に検定で調べ上げる，ということはあまり聞かない．そこで，非常に多くの検定を行って優秀と言われている疑似乱数を棄却する例をできるだけ多く収集するために，このような研究に至ったのである．

1　$\{0,1\}$-疑似乱数の有限次元分布に関する直交検定系

本稿で扱う疑似乱数はすべて $\{0,1\}$ に値を取るもので公平な硬貨投げの確率過程をモデルとする．

有限次元分布を調べるために，以下では一般に $\{0,1\}^k$ に値を取る確率変数 X の分布について考える．任意の $\varepsilon \in \{0,1\}^k$ に対して $P(X=\varepsilon)=2^{-k}$ のとき，X は一様分布して

いるという．X が疑似乱数のとき，一般に一様分布していないかもしれない．そこで一様分布から「ずれ」を検定によって検出したい．

検定法にも様々なタイプが考えられるが，ここでは最も簡単な次のような仮説

$$P(X \in A) = \frac{\#A}{2^k}$$

の検定を考えよう．従って，1 つの検定法は 1 つの部分集合 $A \in \{0,1\}^k$ またはその定義関数によって定まるので，しばしばそれらを同一視する．

定義 1. $\{0,1\}^k$ の部分集合の系 $\{A_i\}_{i=1}^K$ が直交系であるとは，各 A_i が空集合でなくて，かつ

$$\frac{\#(A_i \cap A_j)}{2^k} = \frac{\#A_i}{2^k} \cdot \frac{\#A_j}{2^k}, \qquad i \neq j, \tag{1}$$

が成り立つことをいう．この条件は $\{0,1\}^k$ 上の一様確率測度 μ のもとで，A_i と A_j が独立であることと同値である．

$\{A_i\}_{i=1}^K$ が直交系のとき，各 i について ある確率空間 (Ω, P) 上の $\{0,1\}^k$-値確率変数 Y_i が存在して次を満たす：$P(Y_i \in A_i) = 1$ かつ $P(Y_i \in A_j) = \#A_j/2^k$，$j \neq i$．つまり，$A_i$ でのみ分布の偏りが検出される確率変数が存在する．実際，A_i 上で一様確率測度 μ_{A_i} を考えた確率空間 (A_i, μ_{A_i}) 上の確率変数 $Y_i(x) := x, \quad x \in A_i,$ は上の条件を満たす．

命題 1. $\{A_i\}_{i=1}^K$ が直交系であるための必要十分条件は

$$\mathrm{Cov}\left(\mathbf{1}_{A_i}\mathbf{1}_{A_j}\right) := \mathbf{E}\left[(\mathbf{1}_{A_i} - \mu(A_i))(\mathbf{1}_{A_j} - \mu(A_j))\right] = 0, \qquad i \neq j. \tag{2}$$

ただし，$\mathbf{E}$ は μ による平均を表す．

証明. 容易なので省略. □

命題 2. $\{A_i\}_{i=1}^K$ を $\{0,1\}^k$ の直交検定系とする．任意の $N \in \mathbf{N}$ に対して，$\{0,1\}^k$ の N-直積 $\{0,1\}^{kN}$ を考え，$x \in \{0,1\}^{kN}$ の成分表示を $x = (x_1, \ldots, x_N)$，$x_l \in \{0,1\}^k$，と書く．このとき，任意の $\rho_i > 0$ に対して $\{0,1\}^{kN}$ 上の検定系 $\{\tilde{A}_i\}_{i=1}^K$ を次のように構成する．

$$\tilde{A}_i := \left\{ x = (x_1, \ldots, x_N) \in \{0,1\}^{kN} \;\middle|\; \frac{1}{N}\sum_{l=1}^N \mathbf{1}_{A_i}(x_l) \leq \rho_i \right\}, \qquad i = 1, \ldots, K.$$

このとき，各 $\tilde{A}_i$ が空集合でなければ，$\{\tilde{A}_i\}_{i=1}^K$ は直交系をなす．

証明. $\tilde{\mu}$ を $\{0,1\}^{kN}$ 上の一様確率測度とする．$\tilde{\mu}$ のもとで，$i \neq j$ のとき，$\sum_{l=1}^N \mathbf{1}_{A_i}(x_l)$ と $\sum_{l=1}^N \mathbf{1}_{A_j}(x_l)$ が独立であることを示せばよい．しかし，これは明らかだろう. □

2 パリティ検定系

命題 3. $i=1,\ldots,2^k-1$ に対して,

$$A_i := \left\{ x=(x_1,\ldots,x_k)\in\{0,1\}^k \;\middle|\; \sum_{n=1}^{k} D_n(i)x_n = \text{odd} \right\}$$

とする. ただし, $D_n(i)$ は i の2進数展開の下から n 番目のビット (0 or 1) である. このとき, 以下のことが成り立つ.
(i) $\{A_i\}_{i=1}^{2^k-1}$ は直交検定系である.
(ii) もし $\{0,1\}^k$-値確率変数 X がすべての $i=1,\ldots,2^k-1$ に対して, $P(X\in A_i)=1/2$ を満たせば, X の分布は一様分布である. (これは, 原理的には一様分布しない確率変数はいずれかの A_i による検定で棄却されることを示す.)

証明. 任意の i をとる. $D_n(i)=1$ なる n を 1 つ固定し, 写像 $\phi_n:\{0,1\}^k\to\{0,1\}^k$ を

$$\phi(x_1,\ldots,x_n,\ldots,x_k) := (x_1,\ldots,1-x_n,\ldots,x_k)$$

のように第 n ビットだけを反転させる写像とすれば, ϕ_n は全単射であり, $\phi_n(A_i)=A_i^c$ だから, $\#A_i=\#A_i^c$ が分かる. すなわち, $\#A_i=2^{k-1}$ である.

(i) 直交性, $\#(A_i\cap A_j)=2^{k-2}$, $i\neq j$, を示そう. $i\neq j$ より, $D_n(i)+D_n(j)=1$ となる $n\in\{1,\ldots,k\}$ が存在する. そのような n を一つ固定する. どちらでも同じことだが, $D_n(i)=1$, $D_n(j)=0$ としよう. 先ほどの写像 ϕ_n によって $\phi_n(A_i)=A_i^c$, $\phi_n(A_j)=A_j$ である. 従って, $\phi_n(A_i\cap A_j)=A_i^c\cap A_j$. これより $\#(A_i\cap A_j)=\#(A_i^c\cap A_j)$ であるから, $\#(A_i\cap A_j)=\#A_j/2=2^{k-2}$.

(ii) を示そう. $X=(X_1,\ldots,X_k)$ を $\{0,1\}^k$-値確率変数とする. 任意の $\varepsilon=(\varepsilon_1,\ldots,\varepsilon_k)\in\{0,1\}^k$ をとる. まず, 次の恒等式に注意する.

$$\prod_{n=1}^{k}(1+s_n) = 1+\sum_{i=1}^{2^k-1}\prod_{n=1}^{k} s_n^{D_n(i)}.$$

ここで $s_n^0=1$ としている. この恒等式より,

$$\prod_{n=1}^{k}\left(1+(-1)^{X_n+\varepsilon_n}\right) = 1+\sum_{i=1}^{2^k-1}\prod_{n=1}^{k}(-1)^{D_n(i)(X_n+\varepsilon_n)}. \tag{3}$$

$X_n\neq\varepsilon_n$ となる n があれば $(-1)^{X_n+\varepsilon_n}=-1$ となって左辺の積が 0 になり, すべての n で $X_n=\varepsilon_n$ ならば左辺の積は 2^k となる. これより, (3) の左辺の平均は

$$2^k P(X_n=\varepsilon_n,\ n=1,\ldots,k) \tag{4}$$

に等しい. 一方, (3) の右辺では $X\in A_i$ ならば $\sum_{n=1}^{k} D_n(i)X_n$ =odd なので, 右辺の平均は

$$1+\sum_{i=1}^{2^k}(-1)^{\sum_{n=1}^{k} D_n(i)\varepsilon_n}\left[P(X\in A_i^c)-P(X\in A_i)\right] \tag{5}$$

であることが分かる．(4) と (5) が等しいことより，もし，$P(X \in A_i) = P(X \in A_i^c) = 1/2$ がすべての $i = 1, \ldots, 2^k - 1$ について成り立てば $P(X = \varepsilon) = 2^{-k}$ であることが分かる．従って X は一様分布する． □

上の $\{A_i\}_{i=1}^{2^k}$ を用いた検定を**パリティ検定法**と呼ぶ．すなわち，$P(X \in A_i) = 1/2$ を帰無仮説として検定を行うのである．

注意．区間 $[0,1)$ から $\{0,1\}^\infty$ への2進展開写像によってLebesgue測度は公平な硬貨投げの確率過程の分布を導く．$[0,1)$ 上の Walsh 関数系のレベルセット (値 1 の逆像) は，2進展開写像によって命題 1 の集合系 $\{A_i\}_{i=1}^{2^k}$ に写る．それで，命題 1 の主張は Walsh 関数系の完全性と直交性 (Lebesgue 測度に対する) と同等である．

2.1 不変量 — Parseval の等式

パリティ検定系の理論的な1つの側面を見よう．

命題 4．$\{A_i\}_{i=1}^{2^k-1}$ は命題 3 の集合系とする．疑似乱数 $X = \{X_n\}_{n=1}^\infty$ は m ビットの種を持ち，X の最初の k 項 $\{X_n\}_{n=1}^k$ は，パス空間 $\{0,1\}^k$，$k \geq m$，の中で 2^m 個の点を占有するとする．このとき，次の等式が成り立つ．

$$\sum_{i=1}^{2^k-1} \left| P(X \in A_i) - \frac{1}{2} \right|^2 = \frac{1}{4}\left(2^{k-m} - 1\right). \tag{6}$$

証明．$\{0,1\}^k$ 上の一様確率測度を μ とする．X のパスの $\{0,1\}^k$ の中で占める部分を F とする．仮定より $\#F = 2^m$，従って

$$||F||^2 := \int_{\{0,1\}^k} |1_F(x)|^2 \, \mu(dx) = 2^{m-k}. \tag{7}$$

集合 $A \subset \{0,1\}^k$ に対して関数

$$\chi_{A_i}(x) := \begin{cases} 1, & (x \in A_i) \\ -1, & (x \in A_i^c) \end{cases} \qquad x \in \{0,1\}^k$$

を定義する．$\{A_i\}_{i=1}^{2^k-1}$ が完全直交系だから，関数系 $\{\chi_{A_i}(x)\}_{i=0}^{2^k-1}$ は $L^2(\{0,1\}^k, \mu)$ の完全正規直交系をなす．ただし，$A_0 := \{0,1\}^k$ とする．Parseval の等式により，

$$\begin{aligned} ||F||^2 &= \sum_{i=0}^{2^k-1} \left| \int_{\{0,1\}^k} 1_F(x) \chi_{A_i}(x) \mu(dx) \right|^2 \\ &= \left(\frac{\#F}{2^k} \right)^2 + \sum_{i=1}^{2^k-1} |\mu(F \cap A_i) - \mu(F \cap A^c)|^2 \\ &= 2^{2m-2k} + \sum_{i=1}^{2^k-1} |2\mu(F \cap A_i) - 1|^2 \end{aligned}$$

$$= 2^{2m-2k} + \sum_{i=1}^{2^k-1} \left|2 \cdot 2^{m-k} P(X \in A_i) - 1\right|^2$$

$$= 2^{2m-2k} + 2^{2m-2k+2} \sum_{i=1}^{2^k-1} \left|P(X \in A_i) - \frac{1}{2}\right|^2$$

これを(7)と合わせれば命題の主張(6)を得る. □

命題4ではもちろん, $k > m$ のときに興味がある. 大きな空間 $\{0,1\}^k$ の中にどんなに上手に 2^m 個の点を配置してもいつも等式(6)が成り立つ. たとえば, 次数 m の $GF(2)$-係数原始多項式を基にするM-系列が $k = m+1$ 次元でどのように分布しているか, 考えよう. このとき,

$$\sum_{i=1}^{2^{m+1}-1} \left|P(X \in A_i) - \frac{1}{2}\right|^2 = \frac{1}{4}.$$

となるが, 実はただ1つ A_i が存在して $P(X \in A_i) = 1$ であり, 他の A_j に関しては $P(X \in A_j) = 1/2$ となっている. ここで $X \in A_i$ はそのM-系列 X を定義する漸化式そのものを表す.

(6)からすぐ分かるように,

$$\max_{1 \le i \le 2^k-1} \left|P(X \in A_i) - \frac{1}{2}\right| \ge \frac{1}{2}\sqrt{2^{-m} - 2^{-k}}$$

である.

3 数値実験例

まだ, 十分な数の実験をしていないので現時点では中間報告に過ぎないが, 以下に一つの実験例を示す.

3.1 方法

$\{X_n\}_n$ を31ビットの疑似乱数列とする. U を同じく31ビットの整数とする. このとき,

$$Z_n(U) := \sum_{j=1}^{31} D_j(X_n) D_j(U) \pmod 2, \qquad n = 1, 2, \ldots,$$

とすれば[1], $\{Z_n(U)\}_n$ は硬貨投げの確率過程を模した疑似乱数となっている. そこで, この $\{Z_n(U)\}_n$ に関する検定を行う.

$$S_n^{(10000)}(U) := \sum_{i=1}^{10000} Z_{i+10000(n-1)}(U)$$

[1] $Z_n(U)$ は X_n に U というマスクをかけ, パリティをとったもの.

とおく．$Z_n(U)$ が i.i.d. であると仮定すれば，各 $S_n^{(10000)}(U)$ は平均 5000，分散 2500 の二項分布に従う．*Mathematica* で正確に計算すると，$P(|S_n^{(10000)}(U) - 5000| \leq 100) = 0.9555742$ である．そこで，今度は確率変数

$$T(U) := \#\{1 \leq n \leq 1000\,;\, |S_n^{(10000)}(U) - 5000| > 100\}$$

の平均は $e = 44.4258$ となる．$T(U)$ のサンプルを 1 つ計算するのに 3.1×10^8 ビットの疑似乱数が必要である．

なお，命題 2 によれば，$\{0,1\}^{3.1\times10^8}$ の部分集合たち

$$A(U)_{\geq\rho} := \{x \in \{0,1\}^{3.1\times10^8} \mid \text{サンプル } x \text{ に対して} T(U) \geq \rho\}, \qquad U = 1, \ldots, 2^{31} - 1,$$

および

$$A(U)_{\leq\rho} := \{x \in \{0,1\}^{3.1\times10^8} \mid \text{サンプル } x \text{ に対して} T(U) \leq \rho\}, \qquad U = 1, \ldots, 2^{31} - 1,$$

はそれぞれ直交検定系をなすことが分かる．

$$Q_{\leq\rho} := \#A_{\leq\rho}(U)/2^{3.1\times10^8} \qquad \text{および} \qquad Q_{\geq\rho} := \#A_{\geq\rho}(U)/2^{3.1\times10^8}$$

とおく．これらは U には依存しない．再び *Mathematica* による計算では

ρ	$Q_{\leq\rho}$
22	1.16543×10^{-4}
21	5.42044×10^{-5}
20	2.40724×10^{-5}
19	1.01841×10^{-5}
18	4.09376×10^{-6}

ρ	$Q_{\geq\rho}$
71	9.95016×10^{-5}
72	5.85223×10^{-5}
73	3.39402×10^{-5}
74	1.94112×10^{-5}
75	1.09507×10^{-5}
76	6.09382×10^{-6}

3.2 メルセンヌ・ツイスターに関する実験

メルセンヌ・ツイスターと呼ばれる疑似乱数生成法[3][2]によるサンプルについて調べた．ここでは初期値(624 個の 32 ビット整数)設定のために線形合同法 $y_n := (1664525 \times y_{n-1} + 1) \bmod 2^{32}$ を用いた[3]．すなわち，$y_0 = s$ を種(32 ビット整数)として，上の漸化式により初期値として $y_0, \ldots, y_{623}$ を計算する．これらの初期値をもとに 32 ビット整数列 X_n を の `genrand_31` という関数[4]によって生成する．

[2]日本規格協会(財)によって JIS-Z-9031「ランダム抜取方法」の改正作業が 1998 年に行われた．改正版はまだ発行されていないが，メルセンヌ・ツイスターはその規格の一部として改正版に掲載予定である．

[3]JIS-Z-9031 に掲載予定の初期値の設定法．

[4]JIS-Z-9031 に掲載予定．この関数は 31 ビット整数の疑似乱数を生成するが，内部ではまず 32 ビット整数を作り，最下位ビットを捨てている．

このとき，初期値として $s := 19660809$ を用いて[5]，

$$T(134239667) = 76 \tag{8}$$

を得た．先の表により，$T(U) \geq 76$ となる確率は 6.09382×10^{-6} である．

もし筆者がでたらめに $U = 134239667$ を選んで(8)を得たのなら，これは非常に確率の小さいことが起こったことになる．だが，実際は $U = 134217729$ (16進整数表示では8000001)から $50,000$ 個の U について $T(U)$ を計算して，その途中で(8)を見つけたのである．だから，公平を期すには(8)のような確率の小さい事象がどれくらいの頻度で起こるかを問題にしなければならない．

すべての実験で初期値を $s := 19660809$ とし，とくに $U = 134217729$ から $50,000$ 個の U について偏差が大きい例($T(U) \leq 22$ または $T(U) \geq 71$)を下の表に示した．

U	$T(U)$
134227883	75
134233248	72
134239667	76
134239937	22
134245382	18
134248595	71

U	$T(U)$
134253416	76
134255018	19
134260320	20
134264250	21
134266998	74
134267575	21

上の表では確率 10^{-5} 程度の事象がたびたび起こっていることが分かる．$T(U) \leq 19$ または $T(U) \geq 75$ なる事象は $50,000$ 回の試行で平均的には約 1.05 回現れるが，それが5回も起こっている．$\{T(U)\}_U$ を独立な検定だと見なすと[6]，このような確率はPoisson分布を用いて計算して，およそ

$$1 - \sum_{k=0}^{4} e^{-1.05} \times \frac{1.05^k}{k!} = 0.0045$$

であることが分かる．従って，この実験に使用したメルセンヌ・ツイスターのサンプルは一様に分布しているとは考えにくい．

参考文献

[1] D.E.Knuth, *The Art of Computer Programming*, 2nd ed., Addison-Wesley, (1981), (邦訳)準数値算法/乱数(渋谷政昭訳)，サイエンス社，(1983).

[2] M.Luby, *Pseudorandmness and cryptographic applications*, Princeton Computer Science Notes, Princeton University Press, (1996).

[5] JIS-Z-9031 に掲載予定の初期値.

[6] 実際，$\{Z_n(U)\}_U$ では相関があっても，$\{A(U)_{\geq\rho}\}_U$ や $\{A(U)_{\leq\rho}\}_U$ ではほとんど無相関になってしまう．なお，これらの相関の詳しい評価はまだ行っていない．

[3] M.Matsumoto and T.Nishimura, Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator, *ACM. Trans. Model. Comput. Simul.*, **8-1**, (1998) 3–30.